

Weekly Report

May 5, 2019

1 Work

1. 本周在进行unpair setting下的图片增强，目前还在测试思路的可能性。准备基于第一阶段的训练结果，训练第二阶段的网络。
2. Adversarial Attack基于神经网络验证了想法，下一步将使用字典直接学习对抗样本。
3. 工作时长：工作日每天9个小时，周末共12个小时，共57个小时。

1.1 工作进度

Table 1: 工作进度

项目	进度	截止时间
DRGraph	正在修改代码	
unpair 低光照图片增强	目前初步的实验效果不佳	
NIPS	基于字典学习Adversarial Attack	2019.5.23

2 Paper Reading

2.1 EAD: Elastic-Net Attacks to Deep Neural Networks via Adversarial Examples

以往的攻击方法只对对抗扰动使用了一种loss（如， l_2 ），本文讨论了同时使用 l_1 和 l_2 loss约束情况下的求解方法。

$$\begin{aligned} & \text{minimize}_{\mathbf{x}} \quad c \cdot f(\mathbf{x}, t) + \beta \|\mathbf{x} - \mathbf{x}_0\|_1 + \|\mathbf{x} - \mathbf{x}_0\|_2^2 \\ & \text{subject to} \quad \mathbf{x} \in [0, 1]^p, \end{aligned}$$

Figure 1: #1

2.2 Synthesizing Robust Adversarial Examples

本文的目的在于增加对抗样本的鲁棒性，比如，样本经过旋转之后仍然具有攻击性 $t(x) = Ax + b$ 。那么我们需要在生成样本的过程中，把这些变化加入到目标函数中。

$$\begin{aligned} & \arg \max_{x'} \quad \mathbb{E}_{t \sim T} [\log P(y_t | t(x'))] \\ & \text{subject to} \quad \mathbb{E}_{t \sim T} [d(t(x'), t(x))] < \epsilon \\ & \quad x \in [0, 1]^d \end{aligned}$$

Figure 2: #2

2.3 Adversarial Patch

只要在图片中加一个特定类别的patch就可以使得模型预测出错。基于Synthesizing Robust Adversarial Examples的工作，但是loss的目标只计算到patch。

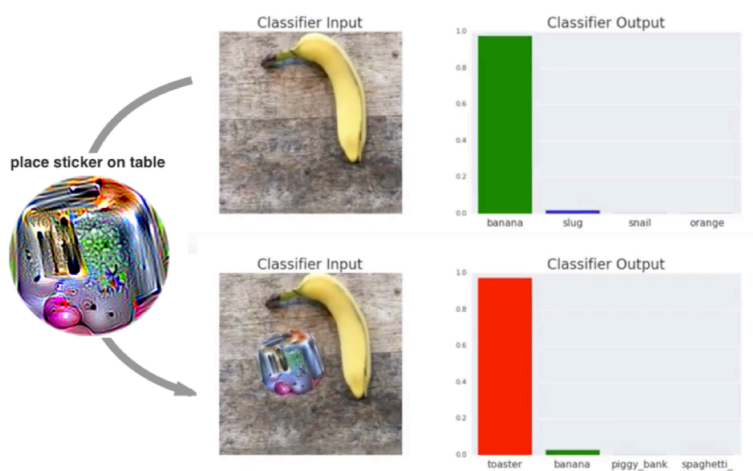


Figure 3: #3

2.4 Robust Physical-World Attacks on Deep Learning Visual Classification

本文考虑了在现实世界中物体的对抗扰动方式，在训练过程中加入了角度、距离和光照等因素，同时还使用了一个mask来保证扰动只修改了目标物体。

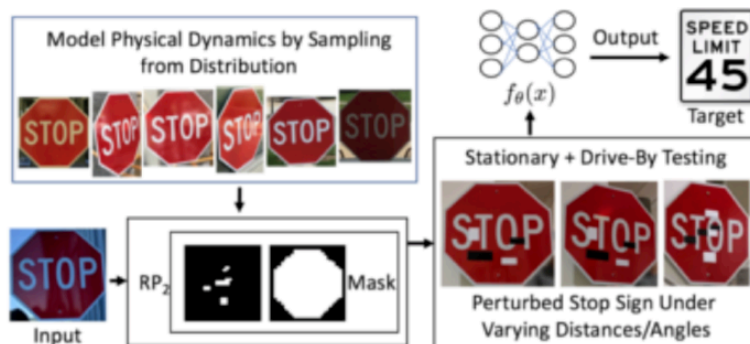


Figure 4: #4